scientific reports

OPEN

Effects and prediction of cognitive load on encoding model of brain response to auditory and linguistic stimuli in educational multimedia

Amir Hosein Asaadi[1,2], S. Hamid Amiri[1], Alireza Bosaghzadeh[1] & Reza Ebrahimpour[3]✉

Multimedia is extensively used for educational purposes. However, certain types of multimedia lack proper design, which could impose a cognitive load on the user. Therefore, it is essential to predict cognitive load and understand how it impairs brain functioning. Participants watched a version of educational multimedia that applied Mayer's principles, followed by a version that did not. Meanwhile, their electroencephalography (EEG) was recorded. Subsequently, they participated in a post-test and completed a self-reported cognitive load questionnaire. The audio envelope and word frequency were extracted from the multimedia, and the temporal response functions (TRFs) were obtained using a linear encoding model. We observed that the behavioral data are different between the two groups and the TRFs of the two multimedia versions were different. We saw changes in the amplitude and latencies of both early and late components. In addition, correlations were found between behavioral data and the amplitude and latencies of TRF components. Cognitive load decreased participants' attention to the multimedia, and semantic processing of words also occurred with a delay and smaller amplitude. Hence, encoding models provide insights into the temporal and spatial mapping of the cognitive load activity, which could help us detect and reduce cognitive load in potential environments such as educational multimedia or simulators for different purposes.

The progress of educational technologies is undeniable; numerous schools now integrate multimedia into classrooms to enhance the learning experience. Making lessons more engaging could help capture learners' attention effortlessly. However, learners may still encounter difficulties in understanding educational multimedia, which could result in increased cognitive load[1]. Cognitive load (CL) can be defined as a multidimensional load, imposed by cognitive tasks on the cognitive system[2]. The aim of research on CL within the instructional domain is to find methods to reduce it and improve the learning process.

CL has been measured with the use of behavioral data[3], secondary tasks[4], eye data[5] and electroencephalography (EEG)[6,7]. Antonekno[8] used frequency bands to measure the CL of a reader. Due to the high temporal resolution of EEG, Event-Related Potentials (ERP) studies can also tell us a lot about the temporal dynamics of the brain activity, particularly by repeatedly presenting conditions time-locked to the stimulus onset. The timing of text-picture integration was explored in ERP and ERD/ERS studies[9]. By plotting the time-frequency of the F_Z and P_Z channels, Scharinger concluded that EEG is a valid and practical tool to measure mental processing demand[10]. In another study, DeLeeuw et al.[11] investigated Mayer's principle of multimedia learning by measuring ERPs using brief visual distractors. Solis[12] utilized ERP and secondary tasks to measure CL while participants were driving in a car simulator. Mobile EEG and Brain Computer Interface (BCI) have also been used to measure CL online[13]. Using ERP analysis, Yu[14] has shown that the degradation of visual stimuli can increase CL. Degraded visual stimuli are intentionally reduced quality or clarity of visual information, such as blurred or pixelated images or text. Sarailoo measured the CL of educational multimedia using machine-learning techniques[15]. For a systematic review, see[16].

[1]Department of Computer Engineering, Shahid Rajaee Teacher Training University, Tehran, Islamic Republic of Iran. [2]Institute for Research in Fundamental Sciences (IPM), School of Cognitive Sciences, Tehran, Iran. [3]Center for Cognitive Science, Institute for Convergence Science and Technology (ICST), Sharif University of Technology, P.O. Box:14588-89694, Tehran, Iran. ✉email: ebrahimpour@sharif.edu

The main issue with the ERP technique lies in its inability to handle complex stimuli, such as those found in educational multimedia. This type of multimedia typically involves the continuous presentation of various elements like pictures, text, animations, speech, and sound. To address this issue and in line with recent studies on the dynamics of brain activity in naturalistic stimuli, we used encoding models. Two of the widely used methods of modeling the brain activity are encoding and decoding[17–20]. Encoding or forward modeling uses stimulus features to predict brain response, while decoding or backward modeling uses brain response to construct stimulus features. Temporal response function (TRF) describes a mapping between some feature of a sensory stimulus and the neural response[21]. TRFs can be used to model the brain activity in different tasks, such as luminance[22], audio envelope[23], low or high-order linguistic features[24,25], second language (L2)[26,27], music[28,29], attention[30–33], and in infants and older adults[34–36]. To the best of our knowledge, no study has investigated the effect of CL on TRFs. To use this encoding model, we employed a continuous stimulus (speech and related visual elements in the form of a series of slides) and recorded a continuous brain response (EEG). Then utilized mTRF toolbox to find TRFs[21,37]. The problem can be addressed using encoding models under two conditions: one with high CL and the other with low CL. Therefore, we analyzed the TRFs of two different educational multimedia to see if the brain responses showed significant differences. For this purpose, we chose two regressors, one for the sensory processing (Audio Envelope), and the other for the higher-order cognition (Word Frequency). The audio envelope describes the changes in sound over time, while word frequency indicates how often a word appears within a corpus. Figure 3a illustrates these regressors.

This study aims to investigate the components of TRF when there is a high CL. The contributions of this work are as follows: (1) Comparing the TRF underlying high vs low CL, and (2) Predicting CL by detecting the relation between components of TRF and behavioral data on the presence of CL. In the section Methods, we describe the experimental design, subjects, and apparatus. Then, in Section "Results", the TRFs, their performance, and their relation to behavioral data are presented. Finally, in Section "Discussion", you can see the interpretation of our data, comparison to previous researches, and limitations.

Methods
Stimulus
From the two listening files, four educational multimedia files were designed. Each listening file contained two multimedia files with different conditions. We used two different listening files to avoid learning in the second session. These multimedia files were designed according to Mayer's principles of multimedia learning. The design violated or applied five extraneous principles of multimedia learning, including: (1) Coherence: remove distracting materials. (2) Signaling: highlighting what to focus on. (3) Redundancy: narration and graphics are better than narration, graphics, and text. (4) Spatial Contiguity: relevant text and visuals are physically close together. (5) Temporal Contiguity: corresponding words and visuals are presented at the same time. Find more details about the stimulus in our previous work[15]. Two linguists in English language teaching devised the scenario for making educational multimedia. Then, the multimedia files were created by a motion graphics specialist in the Adobe After Effects CC 2017 v14.2.1.34 software. The two conditions are with principle (P) and without principle (NP). Figure 1 has three example frames from each multimedia. For each condition, there were two lessons: 11 and 6. The duration of lesson 11 is 342 s, and the duration of lesson 6 is 290 s. The listening files are from Oxford's Open Forum 3[38]; the slides are related images and texts. Each lesson has two versions: P and NP.
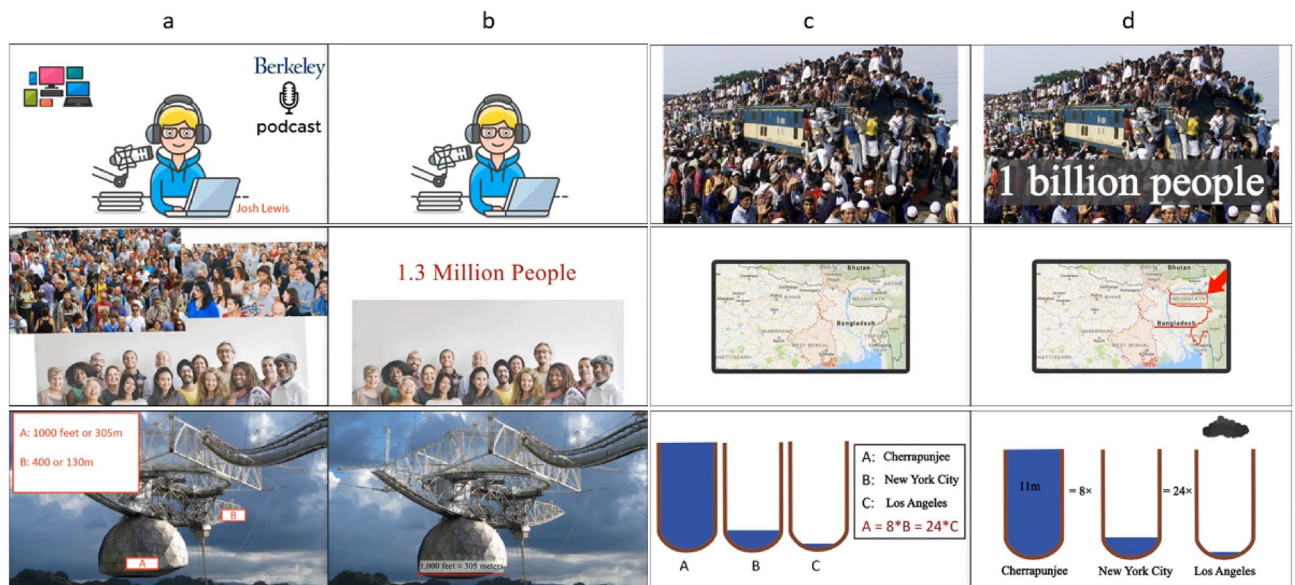
Figure 1. Twelve example frames from designed multimedia. The multimedia is publicly available at https://osf.io/v53np/. (**a**) Three example frames from lesson 11 and without multimedia learning principle (NP) condition. (**b**) The equivalent frames of column (**a**) frames in with multimedia learning principle (P). (**c**) Three example frames from lesson 6 and NP condition. (**d**) The equivalent frames of column (**c**) frames in P condition.

The multimedia and tasks are presented by software designed for this task, available on GitHub[39]. The audio is not changed in any way.

Both P and NP have the same audio, but they have different slides, which differentiates them. In NP multimedia, we violate the principles according to the rules mentioned by Mayer in Multimedia Learning book[40]. The task has only two sessions, one for each multimedia, and it is not divided by trial. The audio sampling rate is 44100 Hz, and the video is 30 frames per second. The video is played on a CRT display monitor (17 inches; PF790; refresh rate, 75 Hz; viewing distance, 57 cm). To minimize head movements and ensure consistent data collection, the subjects put their head on a chin rest.

Participants

Thirty-nine university students, aged between 20 and 29 years (mean = 22.8, std = ±2.5, two females), participated voluntarily after a recruitment announcement. They were randomly assigned to watch lesson 11 NP, then lesson 6 P (n = 21) or lesson 6 NP and then lesson 11 P (n = 18). The data were collected in previous studies[15,41]. Specifically, 29 participants remained in the first condition, and 28 participants remained in the other condition after removing participants due to the following reasons: incomplete recording (n = 2), noisy data (n = 7), to find details see Preprocessing subsection, and too low post-test score (n = 2 in the first condition and n = 1 in the other). All participants reported having normal or corrected-to-normal vision and hearing; None of them had a history of neurological disease.

The first language of all participants is Persian, and their second language is English. They participated in a standard pre-task listening test and accomplished it. In addition, they performed a test similar to the main procedure to familiarize themselves with the main procedure. They signed an informed written consent form before participating in the study. All experimental protocols used in this study were approved by the Iran University of Medical Sciences (IR.IUMS.REC.1397.951). All the methods used in this study were performed in accordance to the guidelines and regulations outlined by the Iran University of Medical Sciences (IR.IUMS.REC.1397.951). The approved protocol was in agreement with the Declaration of Helsinki.

Data collection protocol

The Participant sits in an adjustable chair in a dim light room while no other sounds disturb them. First, the procedure was explained to them, they saw a one-minute video with different subjects from the main video to become familiar with the environment. The EEG recorded using a 32 channel eWave (https://sciencebeam.com) device with a sampling rate of 1 KHz, and the software is eProbe v6.7.3.0, with a cap of 10–20 system[15,41–43]. Two channels are connected to the mastoid bones used for reference. Visual triggers on the monitor were used to ensure synchronization and, there are two loudspeakers in front of the subject, one on the right and one on the left.

As mentioned in the Stimulus section, there are four multimedia, and subjects are randomly assigned to two groups. One group saw Lesson 11 NP and Lesson 6 P, while the other viewed Lesson 6 NP and Lesson 11 P. The experiment consisted of two sessions: first, the eye tracker was calibrated (the eye data is not analyzed in this study, which just served to ensure that subjects were attending to the monitor). Following a countdown, the multimedia presentation began. Then, they are supposed to answer the questions to assess how much they had learned and whether they had paid attention. Subsequently, they completed a paper-based NASA-TLX questionnaire in their first language[44,45]. After a short break, the second session started with the same procedure but with other multimedia and conditions. The entire procedure is illustrated in Fig. 2. Find more details in the Materials and Methods section of our previous work[15].

Analysis

EEG data preprocessing and modeling were performed in MATLAB (MathWorks Inc.), EEGLAB[46], multivariate temporal response function (mTRF) toolbox[21], and modified encoding scripts from the CNSP 2021 workshop (cnspworkshop.net).

Preprocessing
The following steps were undertaken for data preprocessing: First, we filtered the data using both a 1 Hz high-pass filter and a 10 Hz low-pass filter, since neural representation for speech is suggested to occur within this frequency range[47]. To remove noise, we first epoch data to 1 s, then a 200 ms sliding window moved along the EEG data, and whenever the standard deviation of a window exceeds 100, we removed that epoch by replacing it with zero[34]. If a channel has more than 100s, we interpolated that channel, and if a subject has less than 100 s of clear data remaining, that subject is removed from the rest of the analysis. Then we run the Independent Component Analysis (ICA), and to remove non-neural components, we use ICLabel[48]. Then downsampled the data to 250 samples per second and divided it into 10 trials.

Extracting of features from multimedia
Audio Envelope The audio was extracted from the multimedia. To prevent the aliasing effect, we first applied a 1 Hz high-pass filter and a 10 Hz low-pass filter, then downsampled the audio to 250 Hz, similar to the EEG data. Finally, we pick the absolute value of the Hilbert transform as the audio envelope regressor[49].

Word frequency To compute the word frequency regressor, we manually identified the exact timings of the start and end of each word using Praat[50]. Subsequently, we used the SUBTLEX database[51] to determine the
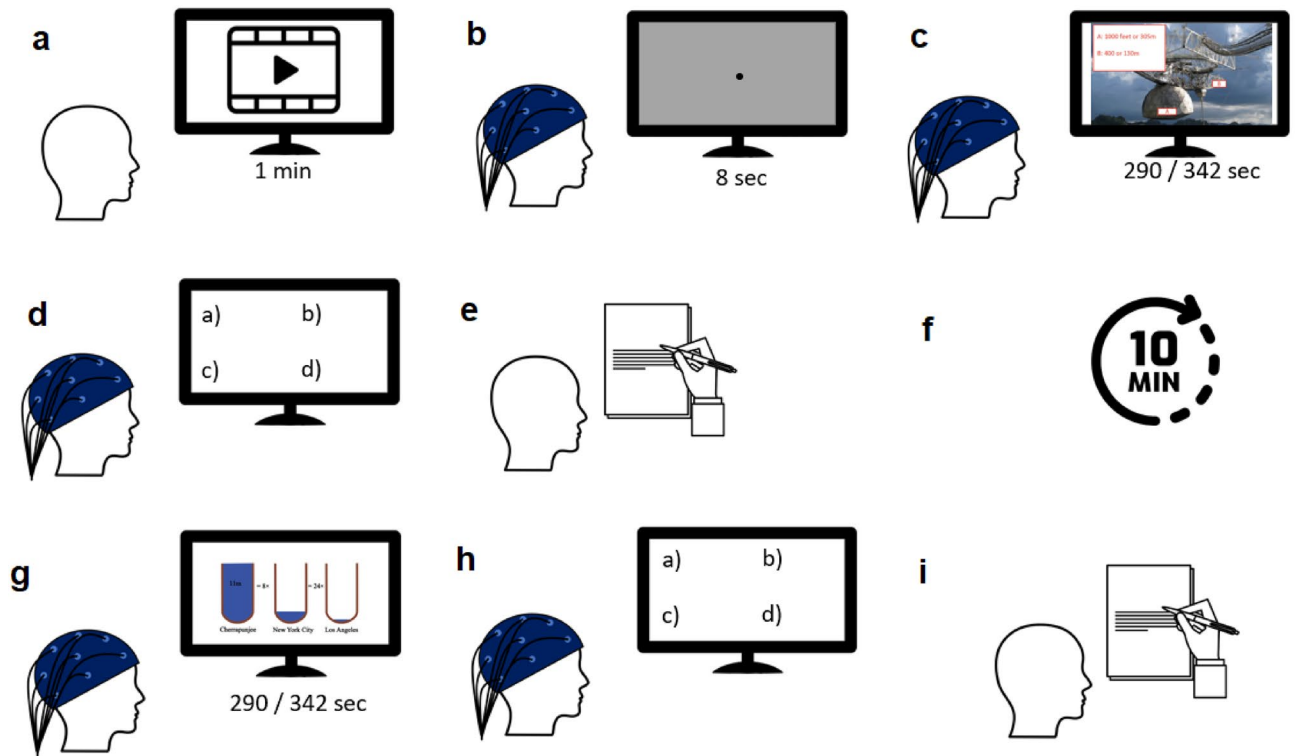
Figure 2. Task Paradigm. Two groups were formed; in one group, Lesson 11 was played first, followed by Lesson 6. In the other group, the sequence was reversed. At the beginning of the experiment, the participants were randomly assigned into one of two groups. (**a**) A one-minute video was shown to the participants to familiarize them with the main experiment. (**b**) Participants were fixated on a point on a gray screen for 8 s to record their baseline EEG. (**c**) While recording their EEG, participants watched Lesson 6 or 11 under the NP condition. (**d**) They attend in a four-answer post-test. (**e**) Participants completed a pen-and-paper NASA-TLX questionnaire. (**f**) They took a ten-minute rest. (**g**) While recording their EEG, participants watched Lesson 6 or 11 under the P condition. (**h**) Again, they engaged in a four-answer post-test based on the most recent multimedia lesson they viewed. (**i**) Finally, they completed another NASA-TLX questionnaire for their last multimedia.

logarithmic value of the word frequency. Here, *wf* stands for the frequency of occurrence of that word in the database, which is calculated based on the subtitles of English movies and TV series. As the highest value in the dataset is 6.329, we set $6.33- \log_{10} wf$ for the duration of each word. In this way, more frequent words are assigned lower values, infrequent words are assigned higher values, and for periods when we do not have a word, we set the value to zero[52]. These features are illustrated in Fig. 3a.

Temporal response function
To model the relationship between the stimulus features and neural data, we use a linear model, specifically ridge regression[53]. The model predicts neural data by a set of time-lagged weights convoluted by input; here is the stimulus feature[22].

$$R(t) = \sum_{1}^{\tau} TRF(\tau)S(t - \tau) + \epsilon(t) \tag{1}$$

In Eq. (1), R is the response at a specific channel at time t, TRF is the temporal response function or model weights, S is the matrix of time-lagged input feature or in other words, design matrix and ϵ is noise or variation that cannot be explained by the model.

To determine the TRF, we can solve an optimization problem by attempting to minimize the error between the predicted and actual responses. If we do so, then the solution is:

$$TRF = \frac{S^T R}{S^T S + \lambda I} \tag{2}$$
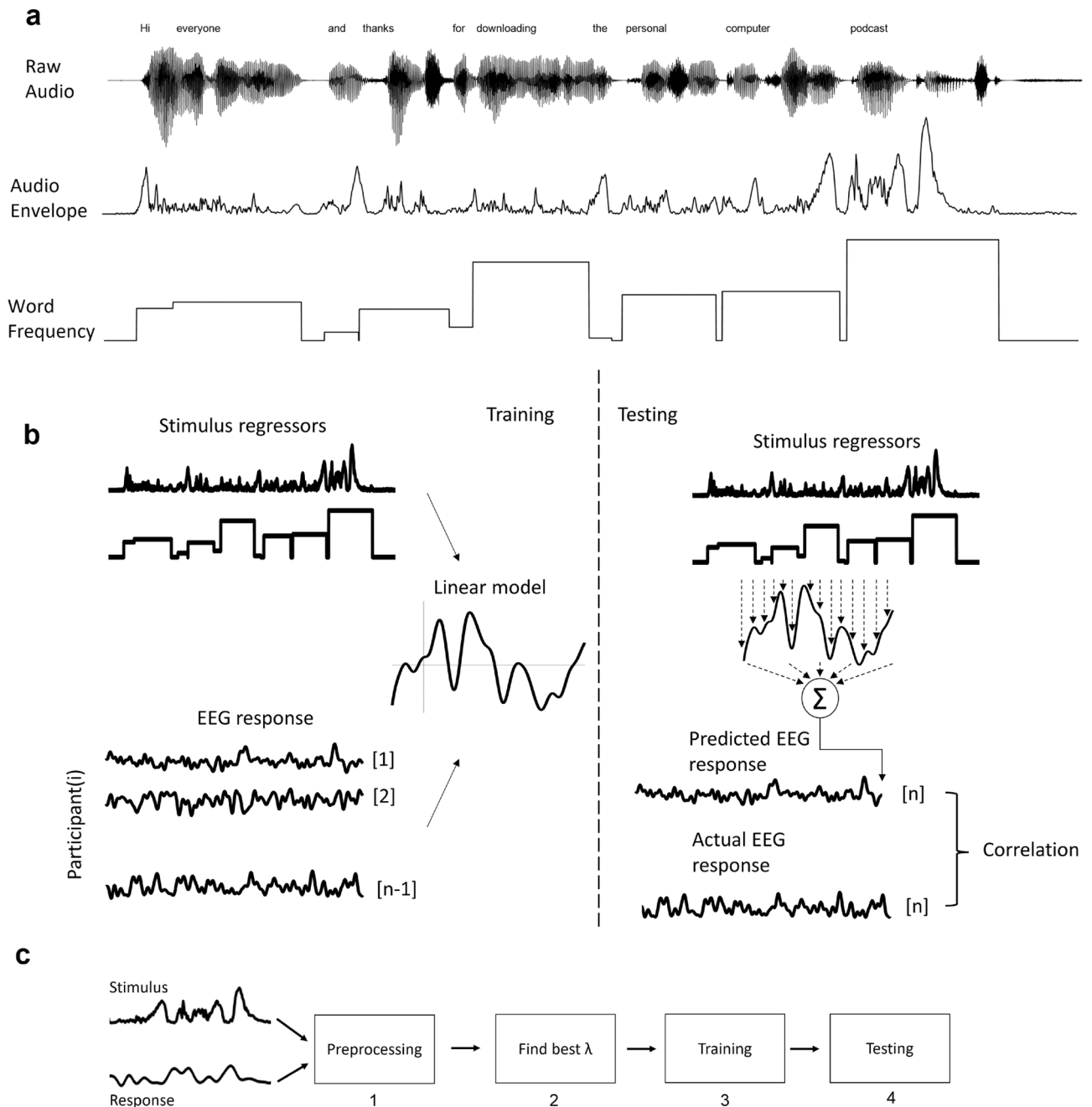
Figure 3. Modeling procedure. (**a**) Regressors extracted from the multimedia are audio envelope and word frequency. (**b**) Generic modeling of continuous EEG data, including the training and testing phase. In training, we find our TRF by optimizing the error of the linear model with n-1 participants and our regressors. Then in the testing phase, predicted response is the convolution of regressors with TRF. Correlation is calculated between actual and predicted EEG responses. (**c**) Procedure of data analysis. (1) Both stimulus and response (EEG) are preprocessed as described in Preprocessing and Extracting of features from multimedia subsection. (2) To find the TRF, first we need to optimize our ridge regression model by finding the best lambda. With a logarithmic vector of lambda values and Eq. (2) we find TRFs for different lambda. Then by Eq. (1), we predict the response for each λ and then calculate the correlation between the predicted and true EEG response. Finally we choose the λ of the highest correlation for the training step. (3) We split data into train and test. In the individual models 80% of each subject and in Generic models n-1 subject. With the best λ found in the previous step and Eq. (2) we find TRF weights. (4) With the test data and founded TRF in the previous step, we calculate the predicted response with Eq. (1). Then, we calculate the correlation of true and predicted EEG response.

In Eq. (2) I is the identity matrix and lambda is a constant value found in the optimization procedure. For the implementation, we use the mTRF Toolbox[21] in MATLAB. First, we divide the continuous data into ten trials. For the τ parameter, we set it between $T_{min} = -200$ ms and $T_{max} = 1000$ ms[34], as numerous previous studies have shown that the processing of auditory information is done in the brain in less than 1 s. Figure 3b shows the modeling procedure.

Choosing the optimal regularization parameter λ
As mentioned in the section Temporal response function, we first equally divide the data into ten trials, then run cross-validation with 80% of the data and use the remaining 20% for testing, ensuring that the model never sees this portion of the data during the training phase. For the training data, we apply the leave-one-out cross-validation method with a range of lambda values to identify the optimal one[54]. For each lambda value, we predict the neural response, calculate the correlation between the predicted and actual data, and select the highest correlation value.

As with EEG data, there are many features and few samples; usually, overfitting is inevitable. To solve this issue, ridge regression uses a penalty parameter named λ. To optimize the model, or in other words, to find the best value for lambda, we create twenty-one lambdas from 10^{-5} to 10^{5} with a 0.5-step size in power[47].

Evaluation of temporal response function
After finding the best value for lambda, we train the model to determine the TRF weights and then test the model on the held-out data. We use correlation to measure how good the prediction of neural response is[31]. As the correlation values are small, we must be sure that they are meaningful. To achieve this, we use a shifted control method.

We create another set of data while the response is circularly shifted by 2 s to ensure that there is no relation between the input stimulus feature and the output neural response. Then, we apply the same modeling process and calculate the model's performance by measuring the correlations[34].

Results

Participants watched two multimedia, P and NP, while their EEG and eye data were recorded. Here, we use only the EEG data in our analysis. After that, to be sure that they attended to multimedia and assess their learning, they participated in the performance test. We also gave them a NASA-TLX questionnaire to ensure that the two versions of the multimedia presentations indeed differ in the levels of CL imposed.

NASA-TLX and post test results

To evaluate whether our designed multimedia in the P and NP conditions affected the subject's CL and check if they were attentive to the multimedia presentations, we asked subjects after watching multimedia to answer 12, four-option questions from multimedia, and then filled out the NASA-TLX questionnaire. We expected to see a significant difference between the P and NP conditions. The results are shown in Fig. 4. As shown, participants who watched the P multimedia reported a lower overall CL compared to those who watched the NP multimedia, and the difference is significant (t(55) = 6.4407403, p = 3.2E−08). The statistical result for post-test also indicates

(a) NASA-TLX scores for two conditions, NP and P

(b) Post test scores for two conditions, NP and P
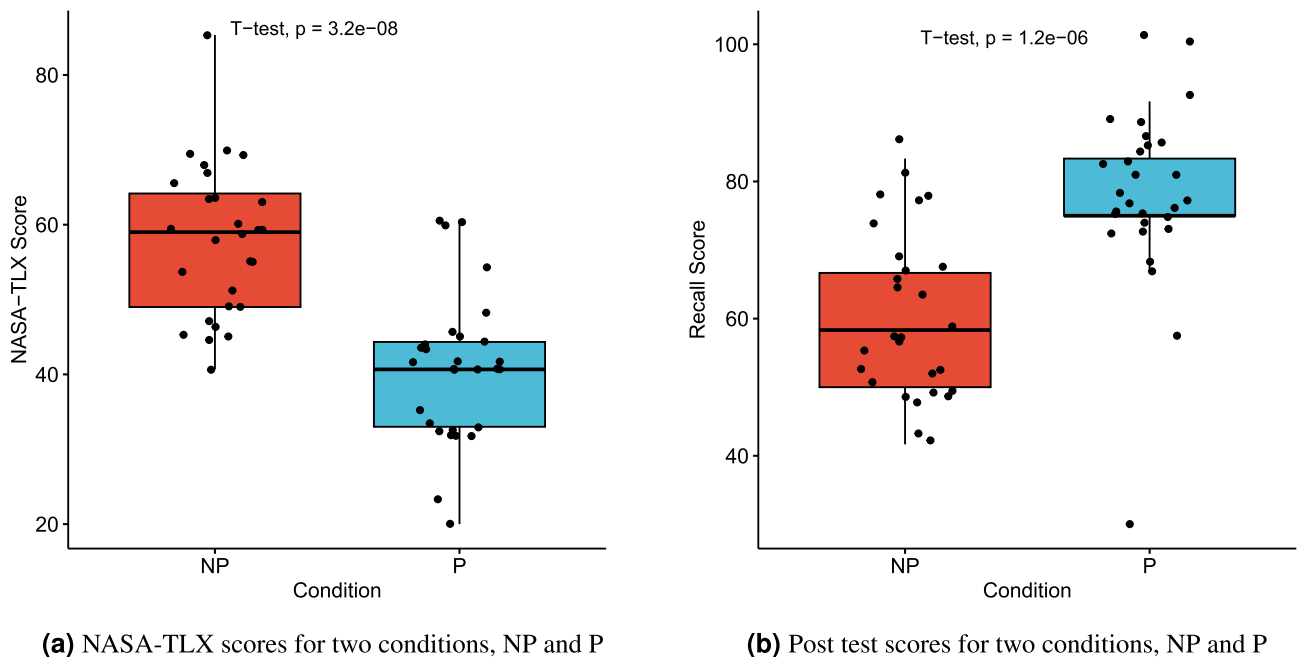
Figure 4. Results of the NASA-TLX and post tests. The scores are in the range of 0 to 100. Each dot represents a subject. The p-value of the t-test is annotated in the figure. (**a**) The results for NASA-TLX task load index and (**b**) The results for our post test.

that the NP multimedia significantly increases CL, and as a result, the post test score for the P condition is higher than that for the NP condition (t(55) = 5.439136037, p = 1.2E−06).
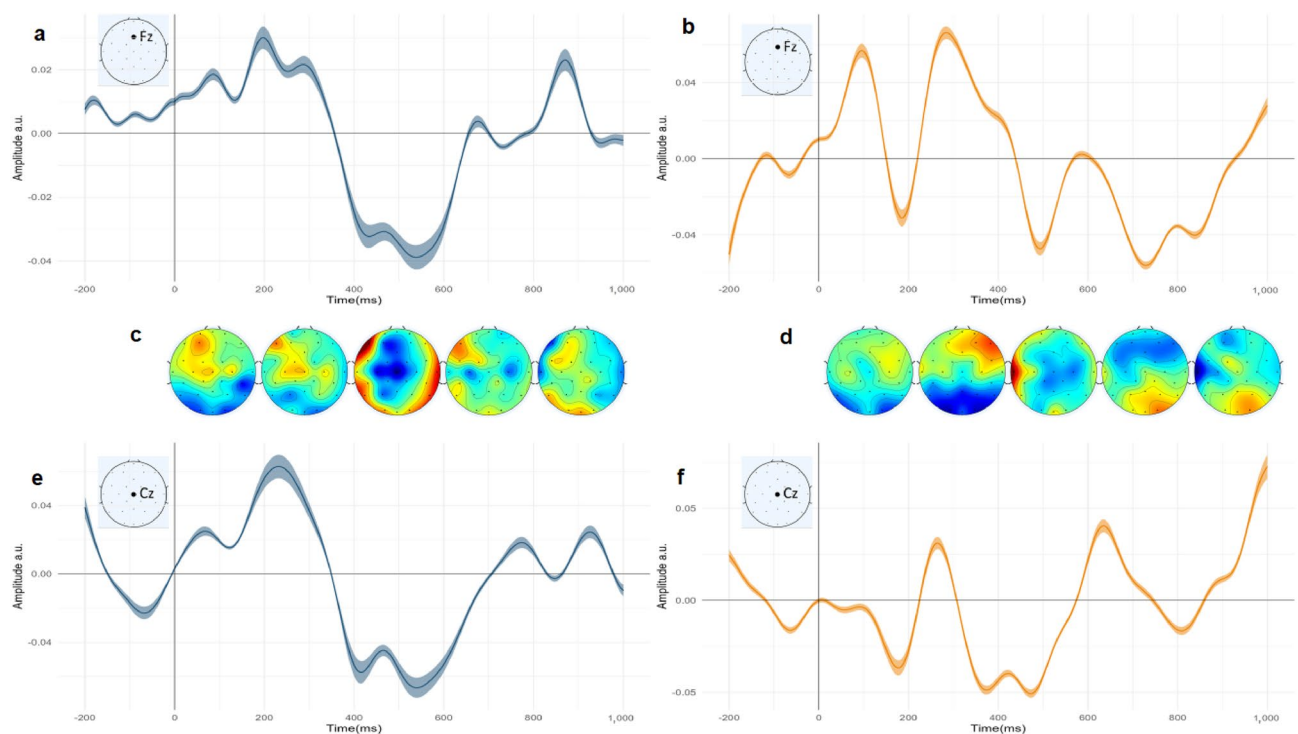
Temporal response function

To see the effect of neural signature in two groups, we try to fit two different regressors, both in the individual and the generic manner in which the model is trained by n-1 subjects and tested by the remaining one (leave-one-out method)[34]. Our regressors were the Audio Envelope and Word frequency. Interestingly, we observed robust, consistent, and different activities in both groups. In the NP group, we saw both early and late neural components, a similar pattern to Jessen et al.[34], and in the P group, as illustrated in Fig. 5, we see two major late components and small early components. In the NP group (channel Fz), the components occur at 100, 200, 300, 500, and 750 ms. In the P group, we observe the components occurring at 400 and 700 ms (n = 29). These TRFs are calculated and plotted for channels Fz and Cz. The average response is shown with a line and the standard error of the mean SEM (95% confidence interval plotted with shadow around the mean). We also plot topographies of brain activities for each 200 ms and put them in a corresponding place on the x-axis. For the P group, we see that around 400 to 600 ms a high activity is happening in the central and frontal regions, with mostly left-lateralized activity. In the NP condition, the duration of activity at the same time intervals was shorter than in the P condition.

For the word frequency regressor, we also observed a robust and different pattern between the two groups. In the P group (channel Fz), Fig. 6 shows three main components occurring at 100, 500, and 850 ms, while in the NP group, we see two main smoother and lower activities. Interestingly, the amplitude in the NP condition was smaller than that in the P condition. The topographies indicate that, similar to the audio envelope regressor, there is a high activity from 400 to 600 ms, but it is happening in frontal electrodes rather than centro frontal electrodes in the auditory P condition. For the NP condition, the activity that is happening is not as strong as that in the P condition.

Generic versus individual response functions

There are two approaches for modeling our data: individual and generic models. In individual models, we train and test our model with one specific subject, whereas in generic models, we train it with n-1 subjects and test it with the remaining subject. When dealing with limited data, generic models are recommended.

In the previous section, we investigated TRFs in a generic manner. Here, we estimate the predictive accuracy of neural data in our forward model. Figure 7 shows the results for both the NP and P groups. As the correlation

Figure 5. Compare Audio envelope temporal response functions of P and NP condition. Standard Error of the Mean (SEM) and 95% confidence interval, as well as topographies of related brain activity over time. The light blue TRF represents the P condition, while the orange one corresponds to the NP group. The corresponding location on the electrode map is illustrated in the corner of each TRF (**a**) Audio envelope (AE) TRF for channel Fz, for P condition. (**b**) AE TRF for channel Fz, for the NP condition. (**c**) Topographies of AE for the P condition. Each topoplot is responsible for the corresponding time in the x-axis of TRF. For example, the first one is from 0 to 200 ms. (**d**) The same topographies for the NP condition. (**e**) AE TRF plotted for channel Cz for the P condition (**f**) AE TRF for the NP condition, channel Cz.
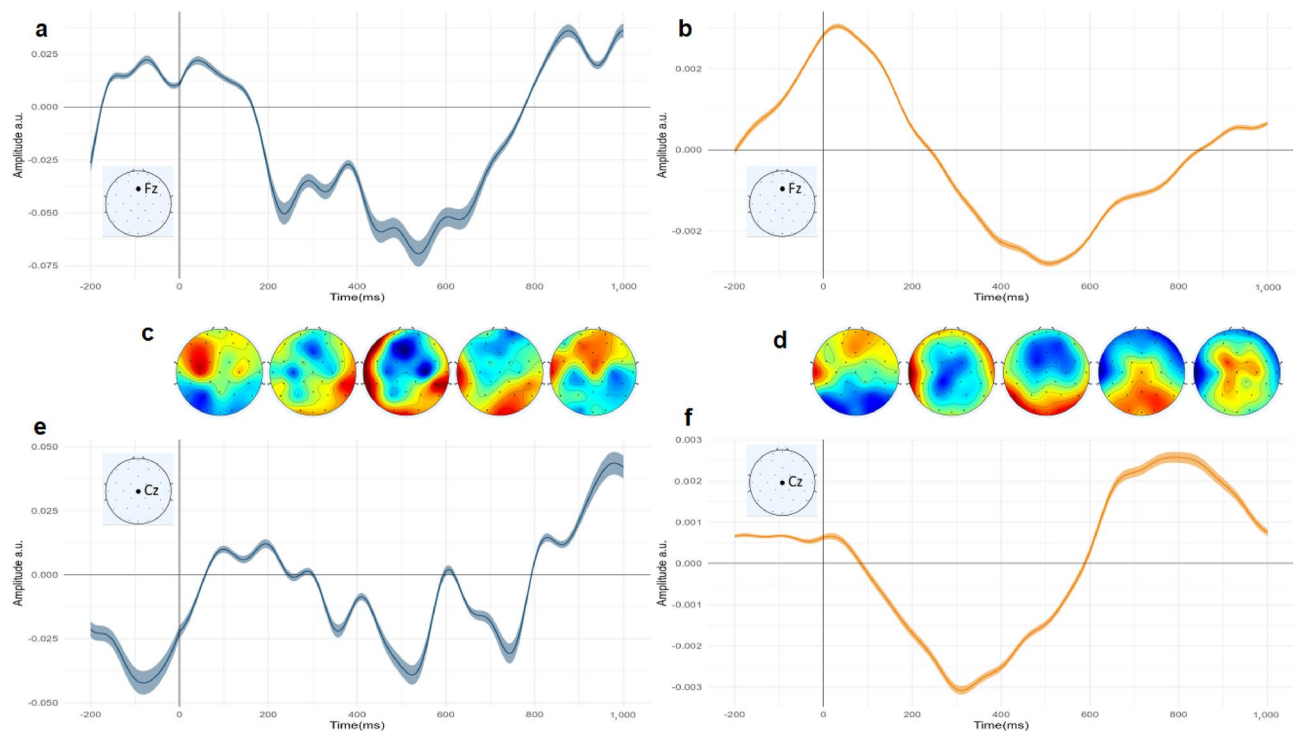
Figure 6. Compare Word Frequency TRFs of P and NP condition. SEM and 95% confidence interval. And topographies of related brain activity across time. The light blue TRF is for P and, the orange one is for the NP group. The corresponding location in the electrode map is illustrated in the corner of each TRF. (**a**) Word frequency (WF) TRF for channel Fz, for P condition. (**b**) WF TRF for channel Fz, for NP. (**c**) Topographies of WF for P condition. Each topoplot is responsible for the corresponding time in the x-axis of TRF. For example, the first one is from 0 to 200 ms. (**d**) Same topographies for the NP condition. (**e**) WF TRF plotted for channel Cz, condition P. (**f**) WF TRF for condition NP, channel Cz.

values are small, we must be sure that they are not just random values, so we shift the EEG signals by 2 s from our regressor signal. Two seconds could be sufficient to ensure that the neural signal is no longer related to the stimulus anymore[55]. Then we generate shifted trials.

First, we see that the mean of the correlations is greater than zero, and they are not close to zero, as in the shifted trials. As expected, individual performance has more variance than generic models. In addition, we observe that the performance for word frequency is higher than that of the audio envelope.

### Relation between features of components and behavioral data

To see if there is any relation between generic TRF components and behavioral data, we divide the time range of TRF into four parts. First from 0 to 150 ms (N/P1), second from 150 to 250 ms (N/P2), third from 250 to 350 ms (N/P3), and the late component for weights found greater than 350 ms (late positivity/negativity). Then, we extract the amplitudes and latencies of each component in the TRFs for the audio envelope and word frequency in the two conditions, NP and P, both in the Cz and Fz channels. We calculate Pearson's correlation and the corresponding p-value with behavioral data, including both the subject's performance on post test and NASA-TLX scores. The results are presented in Table 1 and Fig. 8.

Discussion

In this study, we presented two types of multimedia to participants, P and NP, while simultaneously recording their EEG. The multimedia presentations had the same audio, but the slides were designed in a way to impose high or low CL. We observed that while audio and words had the same content in the two conditions, the brain processed them slightly differently. In addition, we found relationships between the behavioral data and our forward encoding model features. This may help us understand the neural mechanisms underlying high CL. Detecting high CL in ecologically valid setups may help avoid potential losses. It can be beneficial in the following fields: Human-Computer Interaction (HCI), User Experience (UX) Design, Educational technology, Automotive, Aerospace, Industrial engineering, Designing workspaces, and Medical Equipment and Healthcare. Attaining a deeper understanding of the CL would help engineers design systems and environments which use cognitive resources optimally, leading to an enhanced performance, safety, and well-being.

CL can be assessed subjectively, by performance tests and physiological data[56]. Questionnaires like NASA-TLX and performance tests are slow, as their indexes are only available at the end of the task. Thus, for a more comprehensive evaluation of CL over time, we can turn to physiological data such as EEG signals. Previous research indicates that ERP is a suitable method for demonstrating changes in the load on working memory and
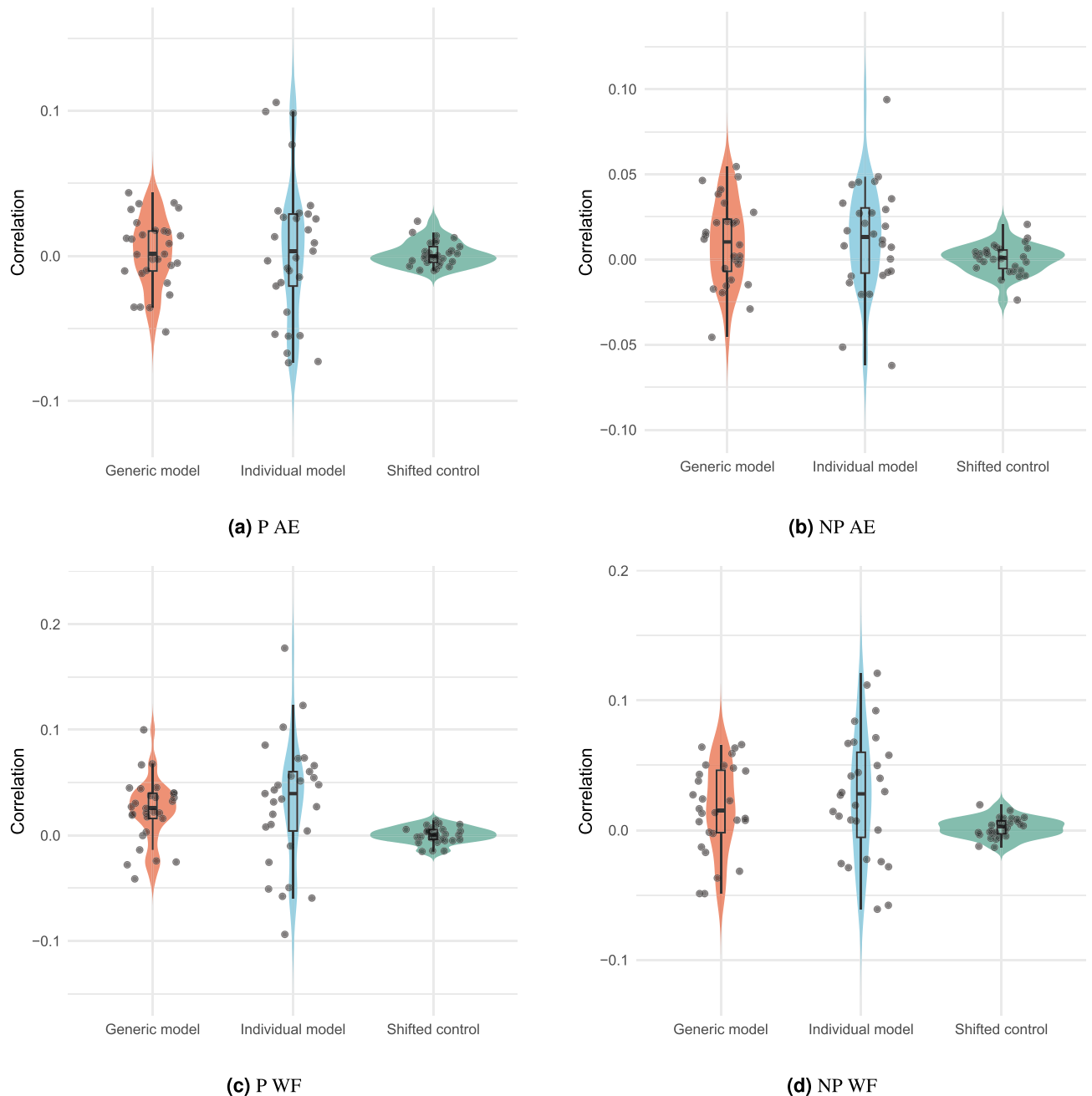
Figure 7. Predictive accuracy of Word Frequency and Audio Envelope for P and NP conditions. The correlation between predicted and actual EEG on 20% of data for two models. Generic (leave-one-out) and Individual models, We also plot the performance for the shifted control condition. All for channel Fz. (**a**) Performance for audio envelope TRF for P condition (**b**) Performance for audio envelope TRF for NP condition (**c**) Performance for Word Frequency TRF for P condition (**d**) Performance for Word Frequency TRF for NP condition.

the various components influenced by different tasks[57]. While ERP provides a high time resolution, TRF enables us to utilize more naturalistic and intricate stimuli commonly encountered in real-world situations. In addition, TRF analysis is effective when we need more engaging stimuli; and so it would be helpful in special groups of subjects such as children and patients.

The early auditory evoked potentials (AEP) are N1 and P2. Several studies show the relation between N1 amplitude and latency with attention. There is a positive correlation between N1 amplitude and attention, that is, the higher the amplitude of N1, the more attention exists[58–60]. Analogous to N1, P2 is modulated by attention, in contrast to N1, it has a negative correlation with attention, in other words, the larger the amplitude of P2 the lower the attention is[61]. Also, it has been shown that P2 amplitude increases during sleep[62]. We find a P1 relation to the NASA-TLX score, which supports the literature[12] indicating that P1 amplitude has a negative correlation with CL. A high load may occur because of Mayer's rules that have been violated, such as extra materials on the slide, the distance between learning objects in the slide, and unsynced audio, all of which can cause a lack of
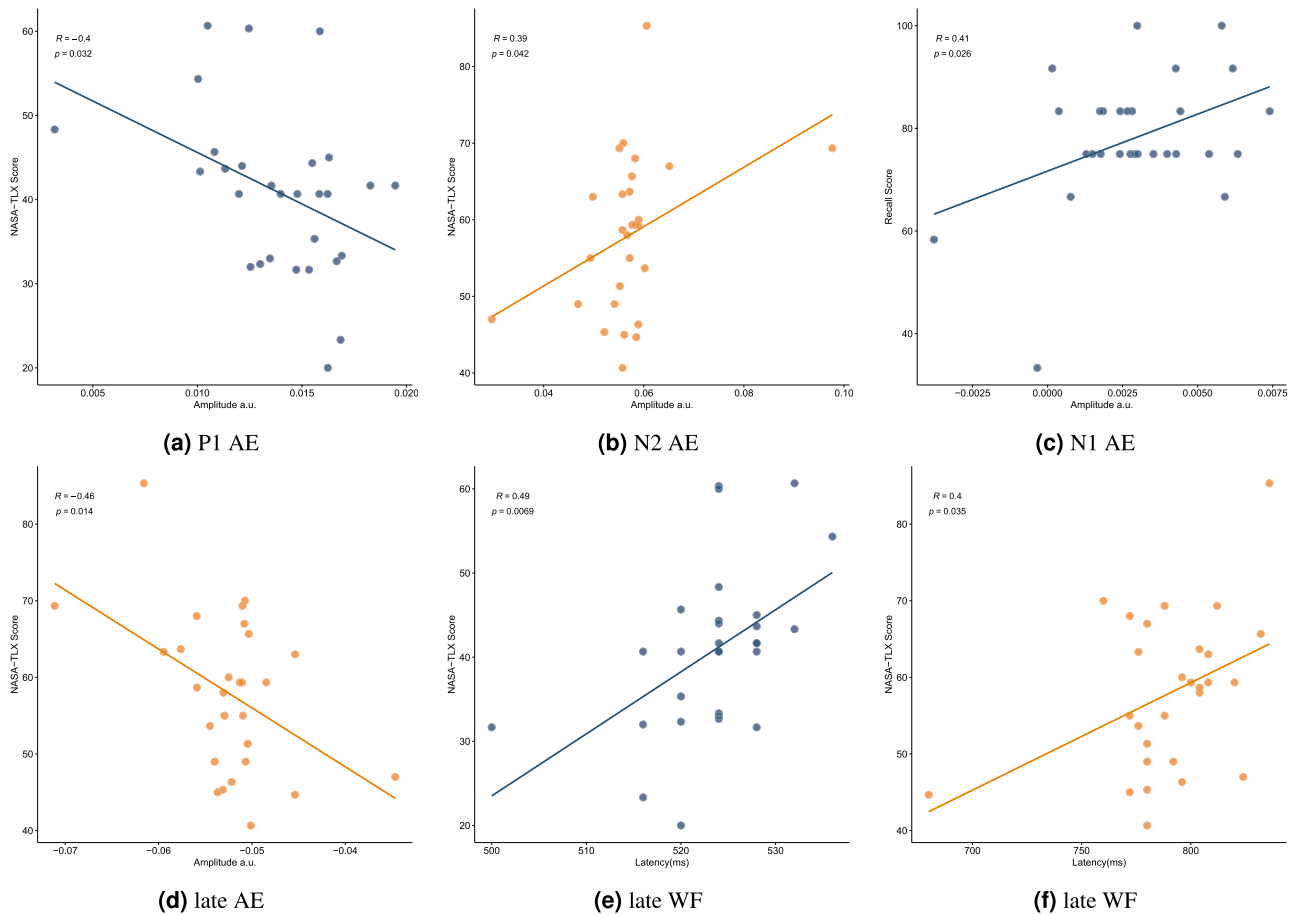
Figure 8. Relation between features of components and behavioral data The relation found for audio envelope (AE) and word frequency (WF) in two conditions P and NP, for electrode Cz and Pz while our behavioral data is NASA-TLX and post test. Each circle represents a subject. The blue color is for the P condition, and the orange is for the NP condition. (**a**) Amplitude of the first component in AE P to NASA-TLX, channel Fz (**b**) Amplitude of the second component in AE NP to post test, channel Fz (**c**) Amplitude of the first component in AE P to NASA-TLX, channel Cz (**d**) Amplitude of late component in AE NP to NASA-TLX, channel Cz (**e**) Latency of late component in WF P to NASA-TLX, channel Cz (**f**) Latency of late component in WF NP to NASA-TLX, channel Cz.

Component	Feature	Ch	Stimulus	Behavioral measure	Condition	r	p values
P1	amp	Fz	Audio Envelope	NASA-TLX	P	-0.400080395	0.031518771
N2	amp	Fz	Audio Envelope	NASA-TLX	NP	0.385990373	0.042484903
N1	amp	Cz	Audio Envelope	post test	P	0.413893684	0.025614097
Late Negativity	amp	Cz	Audio Envelope	NASA-TLX	NP	-0.461061067	0.013535079
Late Negativity	lat	Cz	Word Frequency	NASA-TLX	P	0.490181391	0.006946695
Late Positivity	lat	Cz	Word Frequency	NASA-TLX	NP	0.400161124	0.034859431

Table 1. Correlation between TRF features and behavioral data. The maximum and minimum amplitude of components and the latency of them found for each subject. Then, correlations and p-values between TRF features and the subject's post-test score and NASA-TLX index were calculated and reported. The results show that there is a relation between behavioral data and TRF features.

attention. We also observe that our N1 component has a positive correlation with the post-test score. Previous studies have shown that N1 is responsible for speech segmentation. Detecting the starting point of a word is important for speech comprehension. The earlier and the larger N1, the higher the speech comprehension[63–65]. Ihara et al. also saw that the higher the proficiency in L2, the higher the N1 amplitude[27]. Marcos et al. also see in an ERP study that drivers with a high mental workload show lower N1 amplitude[12]. Previous ERP studies have shown that early components, such as N1 and P2, in audiovisual tasks occur earlier and with lower amplitude[66–68]. This may explain why we see that early components of our TRFs are smaller, and their latency is shorter.

In late components, there is an N400 component. Several studies have shown that N400 is responsible for semantic processing[69,70]. The N400 appears with different modalities, such as audio, words, or pictures. It is suggested that two cognitive processes occur during the N400: first, accessing semantic terms from long-term memory; and second, integrating semantic information together. The amplitude of our late components has a negative correlation with the NASA-TLX score, and the latency of the late components has a positive correlation with the NASA-TLX score. Newman et al. also saw a negative correlation between N400 amplitude and speech comprehension score; we can say that as CL increases, there would be lower space for semantic processing[71]. We also see that there is a positive correlation between the late component latency and the NASA-TLX score. Ihara et al. demonstrated that Japanese individuals with higher L2 proficiency have an earlier N400, indicating that the brain processes the meanings of words earlier than in those with lower proficiencies[27]. Our N400 may occur later under high CL because more time is needed to process semantic information in NP condition.

Figure 8 and Table 1 suggest that CL and learning in educational multimedia can be predicted using merely a small portion of EEG recording, by analyzing the amplitude and latencies of TRF components of only one channel. Friedman et al.[72] also predicted CL in Raven's 2 IQ test, where participants had to solve 36 questions without a time limit while their EEG was recorded. Other studies have used machine learning and statistical methods to measure CL using the same dataset as our research. However, they have not addressed cognitive processes such as audio or linguistic[15,41]. For a review on the CL recognition with EEG see[73].

This study has some potential limitations. Firstly, our focus was on second-language learners. However, to accurately differentiate neural activities related to CL, it would be more appropriate to utilize stimuli in the participants' native language. For instance, even though presenting an excessive amount of text may contradict Mayer's principle, it can assist second-language learners in better understanding multimedia content. Secondly, the hypothesis that every moment in the NP multimedia lacks principles or that each moment in the P multimedia follows principles is probably not correct. For example, in certain moments of NP multimedia, there is both speech and image, these instances could also meet the criteria for the P condition. In addition, by using eye tracking alongside EEG recording, one could also detect where subjects focus their attention. This will allow us to extract a more accurate regressor from the stimuli to use in the forward model[74,75]. Finally, given advancements in technology, particularly in virtual reality (VR) and augmented reality (AR), we suggest examining the effects of these tools on CL[76].

In conclusion, in this study we developed a model to predict CL using EEG signals recorded while participants viewed educational multimedia. The findings indicate that CL influenced the early and N400 components of TRF. Our results highlight the potential to assess CL and learning using TRFs, eliminating the need for traditional paper questionnaires.

Data availability
The raw EEG data and multimedia for TRF analysis are available at https://osf.io/v53np/.

References
1. Mayer, R. E. & Moreno, R. Nine ways to reduce cognitive load in multimedia learning. *Educ. Psychol.* **38**, 43–52. https://doi.org/10.1207/S15326985EP3801_6 (2003).
2. Paas, F. G. W. C. & Van Merriënboer, J. J. G. Instructional control of cognitive load in the training of complex cognitive tasks. *Educ. Psychol. Rev.* **6**, 351–371. https://doi.org/10.1007/BF02213420 (1994).
3. Zhonggen, Y., Ying, Z., Zhichun, Y. & Wentao, C. Student satisfaction, learning outcomes, and cognitive loads with a mobile learning platform. *Comput. Assist. Lang. Learn.* **32**, 323–341. https://doi.org/10.1080/09588221.2018.1517093 (2019).
4. Brünken, R., Plass, J. L. & Leutner, D. Assessment of cognitive load in multimedia learning with dual-task methodology: Auditory load and modality effects. *Instruct. Sci.* **32**, 115–132. https://doi.org/10.1023/B:TRUC.0000021812.96911.c5 (2004).
5. Krejtz, K. Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. *PLoS ONE* **13**, e0203629 (2018).
6. Lin, F.-R & Kao, C.-M. Mental effort detection using EEG data in E-learning contexts. *Comput. Educ.* **122**, 63–79. https://doi.org/10.1016/j.compedu.2018.03.020 (2018).
7. Brouwer, A.-M. *et al.* Estimating workload using eeg spectral power and erps in the n-back task. *J. Neural Eng.* **9**, 045008. https://doi.org/10.1088/1741-2560/9/4/045008 (2012).
8. Antonenko, P., Paas, F., Grabner, R. & van Gog, T. Using electroencephalography to measure cognitive load. *Educ. Psychol. Rev.* **22**, 425–438 (2010).
9. Li, S. *et al.* Dynamic cognitive processes of text-picture integration revealed by event-related potentials. *Brain Res.* **1726**, 146513 (2020).
10. Scharinger, C., Schüler, A. & Gerjets, P. Using eye-tracking and eeg to study the mental processing demands during learning of text-picture combinations. *Int. J. Psychophysiol.* **158**, 201–214 (2020).
11. DeLeeuw, K., Mayer, R. & Giesbrecht, B. When more load leads to less distraction in multimedia learning: An event-related potential approach. In *Proceedings of the Annual Meeting of the Cognitive Science Society* vol. 32 (2010).
12. Solís-Marcos, I. & Kircher, K. Event-related potentials as indices of mental workload while using an in-vehicle information system. *Cogn. Technol. Work* **21**, 55–67. https://doi.org/10.1007/s10111-018-0485-z (2019).
13. Zhou, Y., Xu, T., Cai, Y., Wu, X. & Dong, B. Monitoring cognitive workload in online videos learning through an eeg-based brain-computer interface. In *Learning and Collaboration Technologies. Novel Learning Ecosystems: 4th International Conference, LCT 2017, Held as Part of HCI International 2017, Vancouver, BC, Canada, July 9-14, 2017, Proceedings, Part I 4* 64–73 (Springer, 2017).
14. Yu, K., Prasad, I., Mir, H., Thakor, N. & Al-Nashash, H. Cognitive workload modulation through degraded visual stimuli: A single-trial eeg study. *J. Neural Eng.* **12**, 046020. https://doi.org/10.1088/1741-2560/12/4/046020 (2015).
15. Sarailoo, R., Latifzadeh, K., Amiri, S. H., Bosaghzadeh, A. & Ebrahimpour, R. Assessment of instantaneous cognitive load imposed by educational multimedia using electroencephalography signals. *Front. Neurosci.* **16**, 744737 (2022).
16. Mutlu-Bayraktar, D., Cosgun, V. & Altan, T. Cognitive load in multimedia learning environments: A systematic review. *Comput. Educ.* **141**, 103618. https://doi.org/10.1016/j.compedu.2019.103618 (2019).

17. Fiedler, L. *et al.* Single-channel in-ear-EEG detects the focus of auditory attention to concurrent tone streams and mixed speech. *J. Neural Eng.* **14**, 036020. https://doi.org/10.1088/1741-2552/aa66dd (2017).

18. Wang, L., Wu, E. X. & Chen, F. Eeg-based auditory attention decoding using speech-level-based segmented computational models. *J. Neural Eng.* **18**, 046066. https://doi.org/10.1088/1741-2552/abfeba (2021).

19. Naselaris, T., Kay, K. N., Nishimoto, S. & Gallant, J. L. Encoding and decoding in fMRI. *NeuroImage* **56**, 400–410. https://doi.org/10.1016/j.neuroimage.2010.07.073 (2011).

20. Kriegeskorte, N. Interpreting encoding and decoding models. *Curr. Opin. Neurobiol.* **55**, 167–179 (2019).

21. Crosse, M. J., Di Liberto, G. M., Bednar, A. & Lalor, E. C. The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli. *Front. Hum. Neurosci.* https://doi.org/10.3389/fnhum.2016.00604 *(2016)*.

22. Lalor, E. C., Pearlmutter, B. A., Reilly, R. B., McDarby, G. & Foxe, J. J. The VESPA: A method for the rapid estimation of a visual evoked potential. *NeuroImage* **32**, 1549–1561. https://doi.org/10.1016/j.neuroimage.2006.05.054 (2006).

23. Lalor, E. C. & Foxe, J. J. Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. *Eur. J. Neurosci.* **31**, 189–193. https://doi.org/10.1111/j.1460-9568.2009.07055.x (2010).

24. Di Liberto, G. M., O'Sullivan, J. A. & Lalor, E. C. Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Curr. Biol.* **25**, 2457–2465. https://doi.org/10.1016/j.cub.2015.08.030 (2015).

25. Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J. & Lalor, E. C. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Curr. Biol.* **28**, 803–809. https://doi.org/10.1016/j.cub.2018.01.080 (2018).

26. Liberto, G. M. D. *et al.* Neural representation of linguistic feature hierarchy reflects second-language proficiency. *NeuroImage* **227**, 117586. https://doi.org/10.1016/j.neuroimage.2020.117586 (2021).

27. Ihara, A. S. *et al.* Prediction of second language proficiency based on electroencephalographic signals measured while listening to natural speech. *Front. Hum. Neurosci.* **15**, 665809. https://doi.org/10.3389/fnhum.2021.665809 (2021).

28. Di Liberto, G. M., Pelofi, C., Shamma, S. & de Cheveigné, A. Musical expertise enhances the cortical tracking of the acoustic envelope during naturalistic music listening. *Acoust. Sci. Technol.* **41**, 361–364. https://doi.org/10.1250/ast.41.361 (2020).

29. Di Liberto, G. M. *et al.* Cortical encoding of melodic expectations in human temporal cortex. *eLife* **9**, e51784. https://doi.org/10.7554/eLife.51784 (2020).

30. Power, A. J., Foxe, J. J., Forde, E.-J., Reilly, R. B. & Lalor, E. C. At what time is the cocktail party? A late locus of selective attention to natural speech: A late locus of attention to natural speech. *Eur. J. Neurosci.* **35**, 1497–1503. https://doi.org/10.1111/j.1460-9568.2012.08060.x (2012).

31. Fiedler, L., Wöstmann, M., Herbst, S. K. & Obleser, J. Late cortical tracking of ignored speech facilitates neural selectivity in acoustically challenging conditions. *NeuroImage* **186**, 33–42. https://doi.org/10.1016/j.neuroimage.2018.10.057 (2019).

32. Lesenfants, D. & Francart, T. The interplay of top-down focal attention and the cortical tracking of speech. *Sci. Rep.* **10**, 6922. https://doi.org/10.1038/s41598-020-63587-3 (2020).

33. Paul, B. T., Uzelac, M., Chan, E. & Dimitrijevic, A. Poor early cortical differentiation of speech predicts perceptual difficulties of severely hearing-impaired listeners in multi-talker environments. *Sci. Rep.* **10**, 6141. https://doi.org/10.1038/s41598-020-63103-7 (2020).

34. Jessen, S., Fiedler, L., Münte, T. F. & Obleser, J. Quantifying the individual auditory and visual brain response in 7-month-old infants watching a brief cartoon movie. *NeuroImage* **202**, 116060. https://doi.org/10.1016/j.neuroimage.2019.116060 (2019).

35. Jessen, S., Obleser, J. & Tune, S. Neural tracking in infants: An analytical tool for multisensory social processing in development. *Dev. Cogn. Neurosci.* **52**, 101034. https://doi.org/10.1016/j.dcn.2021.101034 (2021).

36. Broderick, M. P., Di Liberto, G. M., Anderson, A. J., Rofes, A. & Lalor, E. C. Dissociable electrophysiological measures of natural language processing reveal differences in speech comprehension strategy in healthy ageing. *Sci. Rep.* **11**, 4963. https://doi.org/10.1038/s41598-021-84597-9 (2021).

37. Crosse, M. J. *et al.* Linear modeling of neurophysiological responses to speech and other continuous stimuli: Methodological considerations for applied research. *Front. Neurosci.* https://doi.org/10.3389/fnins.2021.705621 *(2021)*.

38. Angela Blackwell, T. N. *Open Forum 3*. https://elt.oup.com/student/openforum/3?cc=ir &selLanguage=en.

39. Latifzadeh, K. *Multimedia Learning hci*. https://github.com/K-Hun/multimedia-learning-hci.

40. Mayer, R. E. *Multimedia Learning* 320. Cambridge University Press.

41. Farkish, A., Bosaghzadeh, A., Amiri, S. H. & Ebrahimpour, R. Evaluating the effects of educational multimedia design principles on cognitive load using EEG signal analysis. *Educ. Inf. Technol.* https://doi.org/10.1007/s10639-022-11283-2 *(2022)*.

42. Vafaei Shooshtari, S., Esmaily Sadrabadi, J., Azizi, Z. & Ebrahimpour, R. Confidence representation of perceptual decision by EEG and Eye data in a random dot motion task. *Neuroscience* **406**, 510–527. https://doi.org/10.1016/j.neuroscience.2019.03.031 (2019).

43. Farzmahdi, A., Fallah, F., Rajimehr, R. & Ebrahimpour, R. Task-dependent neural representations of visual object categories. *Eur. J. Neurosci.* **54**, 6445–6462 (2021).

44. Hart, S. G. & Staveland, L. E. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Adv. Psychol.* **52**, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9 (1988).

45. Mohammadi, M., Mazloumi, A., Naslseraji, J. & Zeraati, H. A. Designing questionnaire of assessing mental workload and determine its validity and reliability among icus nurses in one of the tums's hospitals. *J. School of Public Health Inst. Public Health Res.* **11**, 5043 (2013).

46. Delorme, A. & Makeig, S. EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* **134**, 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009 (2003).

47. Crosse, M. J. *et al.* Linear modeling of neurophysiological responses to speech and other continuous stimuli: Methodological considerations for applied research. *Front. Neurosci.* **15**, 1350. https://doi.org/10.3389/fnins.2021.705621 (2021).

48. Pion-Tonachini, L., Kreutz-Delgado, K. & Makeig, S. ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage* **198**, 181–197. https://doi.org/10.1016/j.neuroimage.2019.05.026 (2019).

49. Yu, W., Sun, W. & Ding, N. Asymmetrical cross-modal influence on neural encoding of auditory and visual features in natural scenes. *NeuroImage* **255**, 119182. https://doi.org/10.1016/j.neuroimage.2022.119182 (2022).

50. Boersma, P. & Weenink, D. *Praat: Doing Phonetics by Computer [Computer Program]*.

51. Brysbaert, M. & New, B. Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behav. Res. Methods* **41**, 977–990. https://doi.org/10.3758/BRM.41.4.977 (2009).

52. Brodbeck, C., Presacco, A. & Simon, J. Z. Neural source dynamics of brain responses to continuous stimuli: Speech processing from acoustics to comprehension. *NeuroImage* **172**, 162–174. https://doi.org/10.1016/j.neuroimage.2018.01.042 (2018).

53. Hoerl, A. E. & Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67. https://doi.org/10.1080/00401706.1970.10488634 (1970).

54. O'Sullivan, A. E., Crosse, M. J., Di Liberto, G. M. & Lalor, E. C. Visual cortical entrainment to motion and categorical speech features during silent lipreading. *Front. Hum. Neurosci.* **10**, 679. https://doi.org/10.3389/fnhum.2016.00679 (2017).

55. Norman-Haignere, S. V. *et al.* A neural population selective for song in human auditory cortex. *Curr. Biol.* https://doi.org/10.1016/j.cub.2022.01.069 *(2022)*.

56. Tao, D. *et al.* A systematic review of physiological measures of mental workload. *Int. J. Environ. Res. Public Health* **16**, 2716 (2019).

57. Ghani, U., Signal, N., Niazi, I. K. & Taylor, D. Erp based measures of cognitive workload: A review. *Neurosci. Biobehav. Rev.* **118**, 18–26 (2020).
58. Hillyard, S. A., Hink, R. F., Schwent, V. L. & Picton, T. W. Electrical signs of selective attention in the human brain. *Science* **182**, 177–180. https://doi.org/10.1126/science.182.4108.177 (1973).
59. Hink, R. F., Hillyard, S. A. & Benson, P. J. Event-related brain potentials and selective attention to acoustic and phonetic cues. *Biol. Psychol.* **6**, 1–16. https://doi.org/10.1016/0301-0511(78)90002-9 (1978).
60. Hansen, J. C. & Hillyard, S. A. Endogenous brain potentials associated with selective auditory attention. *Electroencephalogr. Clin. Neurophysiol.* **49**, 277–290. https://doi.org/10.1016/0013-4694(80)90222-9 (1980).
61. Crowley, K. E. & Colrain, I. M. A review of the evidence for P2 being an independent component process: Age, sleep and modality. *Clin. Neurophysiol.* **115**, 732–744. https://doi.org/10.1016/j.clinph.2003.11.021 (2004).
62. Nielsen-Bohlman, L., Knight, R. T., Woods, D. L. & Woodward, K. Differential auditory processing continues during sleep. *Electroencephalogr. Clin. Neurophysiol.* **79**, 281–290. https://doi.org/10.1016/0013-4694(91)90124-M (1999).
63. Sanders, L. D., Newport, E. L. & Neville, H. J. Segmenting nonsense: An event-related potential index of perceived onsets in continuous speech. *Nat. Neurosci.* **5**, 700–703. https://doi.org/10.1038/nn873 (2002).
64. Sanders, L. D. & Neville, H. J. An ERP study of continuous speech processing: I. Segmentation, semantics, and syntax in native speakers. *Cogn. Brain Res.* **15**, 228–240. https://doi.org/10.1016/S0926-6410(02)00195-7 (2003).
65. Abla, D., Katahira, K. & Okanoya, K. On-line assessment of statistical learning by event-related potentials. *J. Cogn. Neurosci.* **20**, 952–964. https://doi.org/10.1162/jocn.2008.20058 (2008).
66. Sorati, M. & Behne, D. M. Musical expertise affects audiovisual speech perception: Findings from event-related potentials and inter-trial phase coherence. *Front. Psychol.* **10**, 2562 (2019).
67. Ganesh, A. C., Berthommier, F., Vilain, C., Sato, M. & Schwartz, J.-L. A possible neurophysiological correlate of audiovisual binding and unbinding in speech perception. *Front. Psychol.* **5**, 1340 (2014).
68. van Wassenhove, V., Grant, K. W. & Poeppel, D. Visual speech speeds up the neural processing of auditory speech. *Proc. Natl. Acad. Sci. USA* **102**, 1181–1186. https://doi.org/10.1073/pnas.0408949102 (2005).
69. Lau, E. F., Phillips, C. & Poeppel, D. A cortical network for semantics: (De)constructing the N400. *Nat. Rev. Neurosci.* **9**, 920–933. https://doi.org/10.1038/nrn2532 (2008).
70. Kutas, M. & Federmeier, K. D. Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annu. Rev. Psychol.* **62**, 621–647. https://doi.org/10.1146/annurev.psych.093008.131123 (2011).
71. Newman, A. J., Tremblay, A., Nichols, E. S., Neville, H. J. & Ullman, M. T. The influence of language proficiency on lexical semantic processing in native and late learners of English. *J. Cogn. Neurosci.* **24**, 1205–1223. https://doi.org/10.1162/jocn_a_00143 (2012).
72. Friedman, N., Fekete, T., Gal, K. & Shriki, O. EEG-based prediction of cognitive load in intelligence tests. *Front. Hum. Neurosci.* **13**, 191. https://doi.org/10.3389/fnhum.2019.00191 (2019).
73. Zhou, Y. *et al.* Cognitive workload recognition using eeg signals and machine learning: A review. *IEEE Trans. Cogn. Dev. Syst.* https://doi.org/10.1109/TCDS.2021.3090217 (2021).
74. Ali, Q. *et al.* Eye-tracking technologies supporting vision screening in children. In *2020 11th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* 000471–000478. https://doi.org/10.1109/CogInfoCom50765.2020.9237913 (2020).
75. Katona, J. Measuring cognition load using eye-tracking parameters based on algorithm description tools. *Sensors* **22**, 912 (2022).
76. Katona, J. A review of human-computer interaction and virtual reality research fields in cognitive infocommunications. *Appl. Sci.* **11**, 2646 (2021).

Acknowledgements

Author contributions

R.E. Conceived the presented idea, planned the experiment, interpretation, and discussed the results. S.A. and A.B. determined the methodology, including signal processing methods and machine learning approaches. A.A. analyzed the data, discussed the results, and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to R.E.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.